



## Microarray data analysis – two approaches

**Piwowar M\*, Kułaga T\*\*, Jadczyk T\*\*, Malawski M\*\*\*\*,  
Matczyńska E\*, Piątkowska W\*\*\*\*, Roterman I\***

- \* Jagiellonian University – Medical College, Cracow, Poland
- \*\* Jagiellonian University The Faculty of Mathematics and Computer Science
- \*\*\* Academic Computer Centre – CYFRONET
- \*\*\*\* University of Science and Technology, Institute of Computer Science, AGH
- \*\*\*\*\* University of Science and Technology, Faculty of Physics, Astronomy

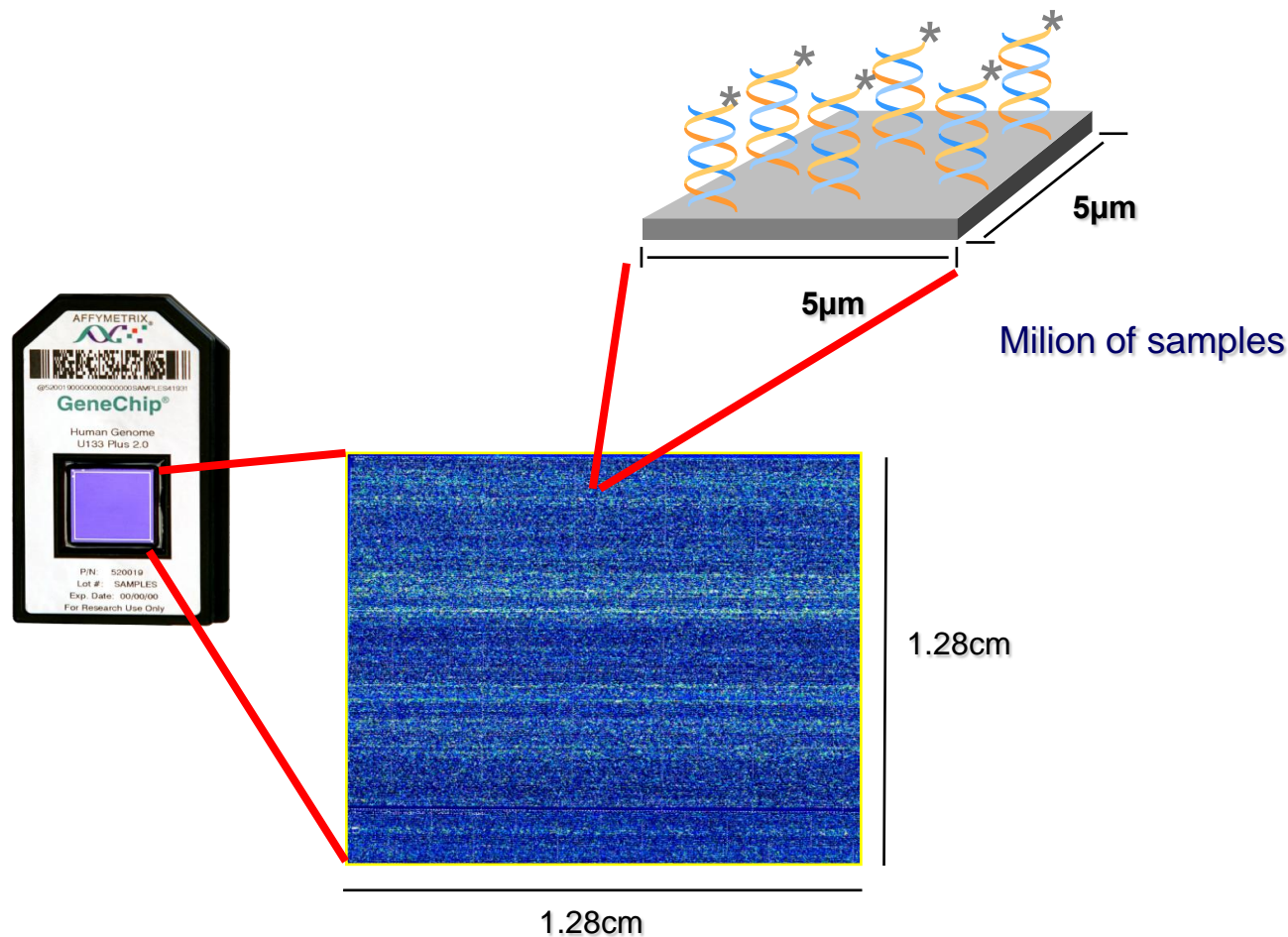


AGH



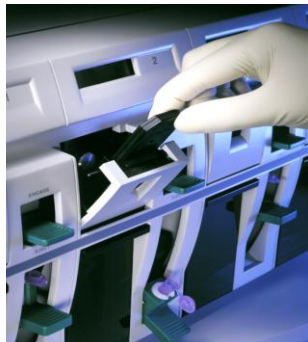


1. **Microarray technique - introduction**
2. **Structure of experiment**
3. **Two methods of microarray analysis**
  - **Measurement of association: genes-biological issue**
  - **Bayesian Networks procedure for creation of genes-biological issue dependence networks**
4. **Virolab – environment for microarray analysis**





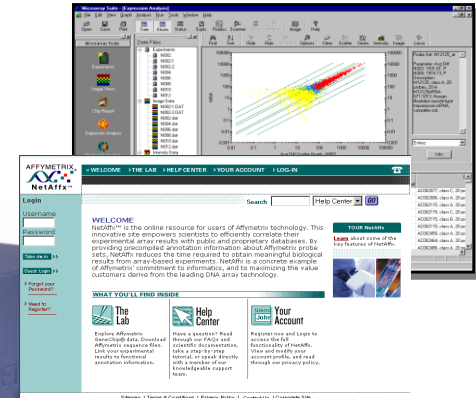
Microarray  
(chips)



Fluidics  
Station



Scanner



Analysis



**Microarrays can be used:**

- **to measure changes in expression levels**
- **to detect single nucleotide polymorphism (SNPs)**
- **in genotyping**
- **in resequencing mutant genomes**




Microarrays can be used:

- to measure changes in expression levels
- to detect single nucleotide polymorphism (SNPs)
- in genotyping
- in resequencing mutant genomes

It helps:

- to discover and understand disease pathways
- to develop better methods of detection, treatment, and prevention





**NCBI**

**National Center for Biotechnology Information**

[National Library of Medicine](#)      [National Institutes of Health](#)

PubMed
All Databases
BLAST
OMIM
Books
TaxBrowser
Structure

Search  for

**SITE MAP**

Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to  
NCBI

**GenBank**  
Sequence  
submission support  
and software

**Literature  
databases**

**▶ What does NCBI do?**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More about NCBI...](#)

**Genome Reference Consortium**

**Hot Spots**

- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools



NCBI  
PubMed All Data  
Search All Databases

**SITE MAP**  
Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to NCBI

**GenBank**  
Sequence submission support and software

**Literature databases**

National Center for Biotechnology Information

**NCBI**

**GEO**  
Gene Expression Omnibus

HOME SEARCH SITE MAP Handout NAR 2006 Paper NAR 2002 Paper FAQ MIAME Email GEO

NCBI > GEO [?](#) Not logged in | [Login](#) [?](#)

**Gene Expression Omnibus:** a gene expression/molecular abundance repository supporting [MIAME compliant](#) data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

**GEO navigation**

QUERY

- DataSets  [GO](#)
- Gene profiles  [GO](#)
- GEO accession  [GO](#)
- GEO BLAST

BROWSE

- DataSets
- GEO accessions
  - Platforms
  - Samples
  - Series

**Public data**

GPL Platforms	5669
GSM Samples	288054
GSE Series	11360
<i>Total</i>	<i>305083</i>

**Site contents**

**Documentation**

Overview | [FAQ](#) | [Find Submission guide](#)  
[Linking & citing](#)  
[Journal citations](#)  
[Programmatic access](#)  
[DataSet clusters](#)  
[GEO announce list](#)  
[Data disclaimer](#)  
[GEO staff](#)

**Query & Browse** [?](#)

[Repository browser](#)  
[Submitter contacts](#)  
[SAGEmap](#)  
[FTP site](#)





**NCBI National Center for Biotechnology Information**  
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search All Databases for Go

**SITE MAP**  
Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to NCBI

**GenBank**  
Sequence submission support and software

**Literature databases**

**What does NCBI do?** Hot Spots

Established in 1988 as a national resource for molecular biology information, NCBI creates Clusters of orthologous groups

**NCBI GEO** Gene Expression Omnibus

HOME SEARCH SITE MAP Handout NAR 2006 Paper NAR 2002 Paper FAQ MIAME Email GEO

NCBI - GEO Not logged in | Login

**Gene Expression Omnibus:** a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

**Public data**  
GPL Platforms 5669  
GSM Samples 288054  
GSE Series 11360  
Total 305083

**GEO navigation**  
DataSets go

**Site contents**  
Join  
FAQ | Find guide  
ing  
ions  
tic access  
sters  
ice list  
ner  
wse  
rowser  
contacts

**Department of Bioinformatics and Telemedicine**  
**Jagiellonian University, Medical College**

**bit**  
12-13.03.2009



**National Center for Biotechnology Information**  
National Library of Medicine | National Institutes of Health

PubMed | All Databases | BLAST | OMIM | Books | TaxBrowser | Structure

Search: All Databases for [ ] Go

**SITE MAP**  
Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to NCBI

**GenBank**  
Sequence submission support and software

**Literature databases**

**What does NCBI do?**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More about NCBI...](#)

**Hot Spots**

- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools

**Genome Reference Consortium**



**Gene Expression Omnibus**

HOME | SEARCH | SITE MAP | Handout | NAR 2006 Paper | NAR 2002 Paper | FAQ | MIAME | Email GEO

NCBI - GEO | Not logged in | Login

**Gene Expression Omnibus:** a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

**Public data**

GPL Platforms	5669
GSM Samples	288054
GSE Series	11360
<b>Total</b>	<b>305083</b>

**Site contents**

**Documentation**  
Overview | FAQ | Find Submission guide  
Linking & citing  
Journal citations  
Programmatic access  
DataSet clusters  
GEO announce list  
Data disclaimer  
GEO staff

**Query & Browse**

Repository browser  
Submitter contacts  
SAGEmap  
FTP site

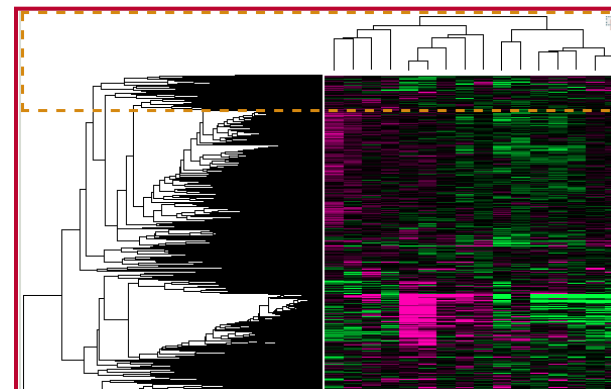
**GEO navigation**

**QUERY**

- DataSets [ ] GO
- Gene profiles [ ] GO
- GEO accession [ ] GO
- GEO BLAST

**BROWSE**

- DataSets [ ] Platforms [ ]
- GEO accessions [ ] Samples [ ]
- [ ] Series [ ]



### Gene Expression Omnibus: a gene expression/molecular abundance repository

GPL	Platforms	<a href="#">5669</a>
GSM	Samples	<a href="#">288095</a>
GSE	Series	<a href="#">11372</a>
<b>Total</b>		<b><a href="#">305136</a></b>

<http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html>



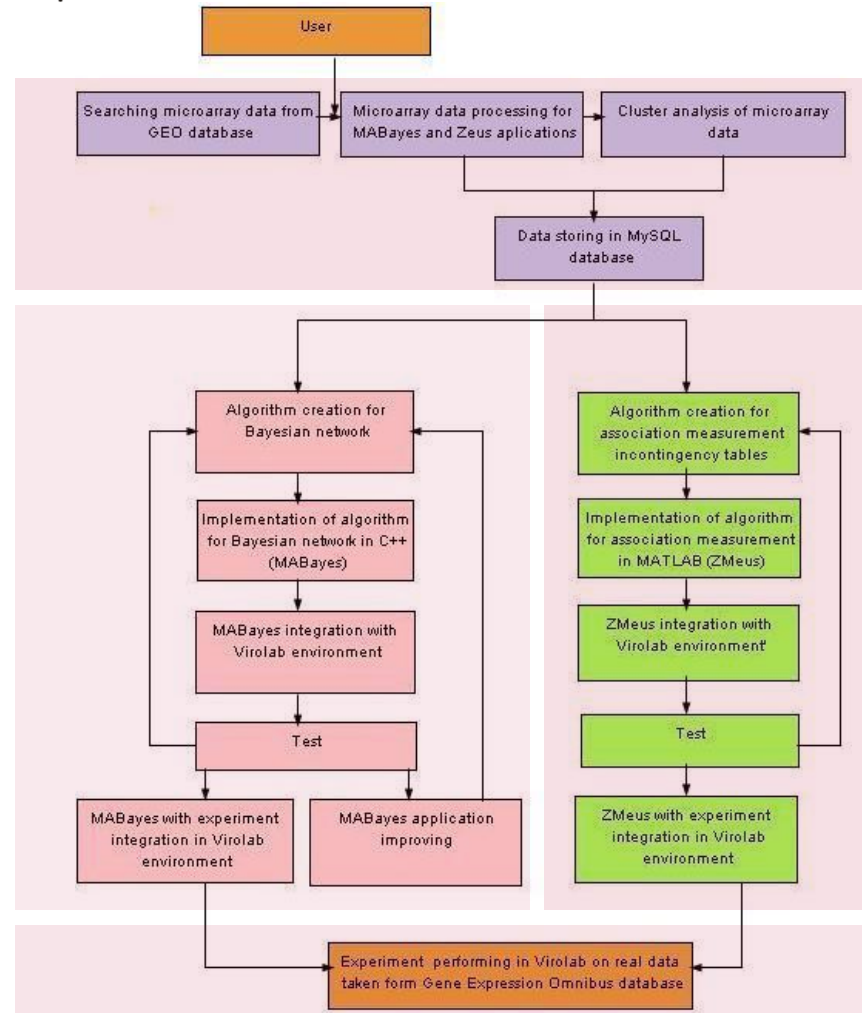
Structure of experiment

Data processing

Z correlation coefficient measurement

Bayesian Networks procedure

Results interpretation





**Gene-biological issue association  
by Z-association coefficient measurement**

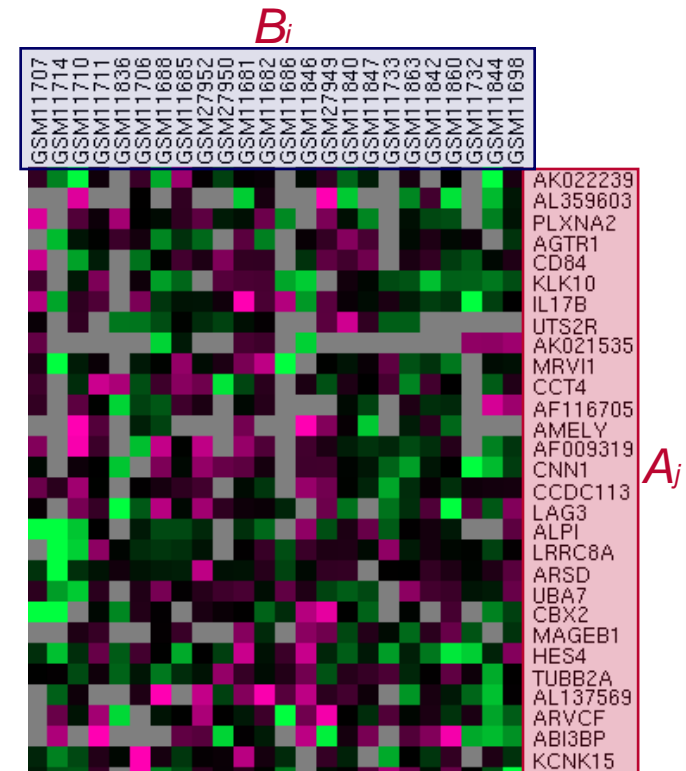


The method is based on a definition of dependence between two families of events:

$A_j$  and  $B_i$

$A_j$  : genes or group of genes (clusters)

$B_i$  : biological issue  
eg. Process, Component, Function

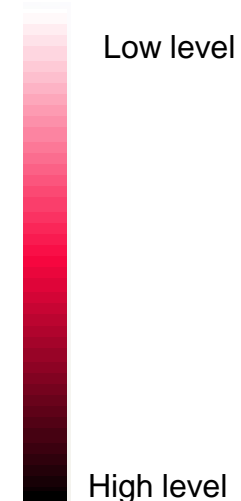




The coefficient is defined on the basis of the probabilities of dependent events. It gives information about dependence between particular genes and eg. biological issue.

$$Z^2(A : B) := 1 - [P(B) \frac{1 - P(A|B)}{P(\bar{A})} \cdot \frac{1 - P(\bar{A}|B)}{P(A)} + P(\bar{B}) \frac{1 - P(A|\bar{B})}{P(\bar{A})} \cdot \frac{1 - P(\bar{A}|\bar{B})}{P(A)}]$$

B→	B <sub>1</sub>	B <sub>2</sub>	...	B <sub>c</sub>	Total
A↓					
A <sub>1</sub>					<i>p</i> <sub>1</sub>
A <sub>2</sub>					<i>p</i> <sub>2</sub>
⋮					⋮
A <sub>r</sub>					<i>p</i> <sub>r</sub>
Total	<i>p</i> <sub>1</sub>	<i>p</i> <sub>2</sub>	...	<i>p</i> <sub>c</sub>	1

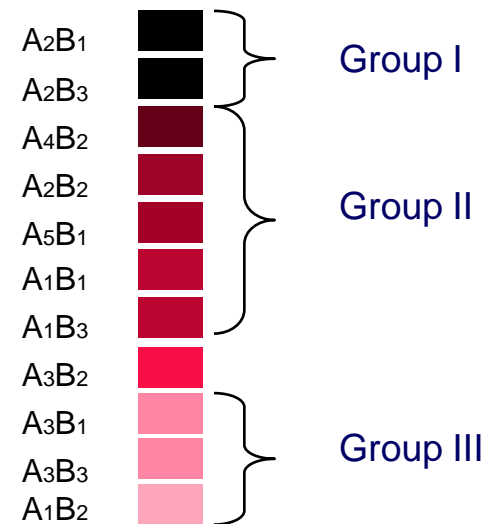




The coefficient is defined on the basis of the probabilities of dependent events. It gives information about dependence between particular genes and biological issue.

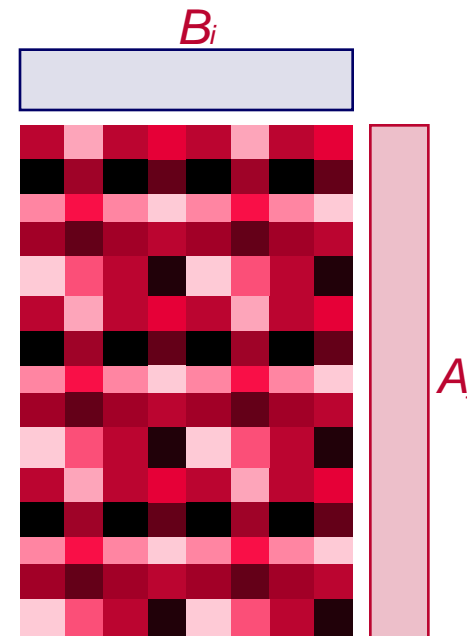
$$Z^2(A : B) := 1 - [P(B) \frac{1 - P(A|B)}{P(\bar{A})} \cdot \frac{1 - P(\bar{A}|B)}{P(A)} + P(\bar{B}) \frac{1 - P(A|\bar{B})}{P(\bar{A})} \cdot \frac{1 - P(\bar{A}|\bar{B})}{P(A)}]$$

B→ A↓	B <sub>1</sub>	B <sub>2</sub>	...	B <sub>c</sub>	Total
A <sub>1</sub>					p <sub>1</sub>
A <sub>2</sub>					p <sub>2</sub>
⋮					⋮
A <sub>r</sub>					p <sub>r</sub>
Total	p <sub>1</sub>	p <sub>2</sub>	...	p <sub>c</sub>	1





1.  $Z(A : B_1; B_2; \dots ; B_n)$
2.  $Z(A_1; A_2; \dots ; A_k : B)$
3.  $Z(A_1; A_2; \dots ; A_k : B_1; B_2; \dots ; B_n)$



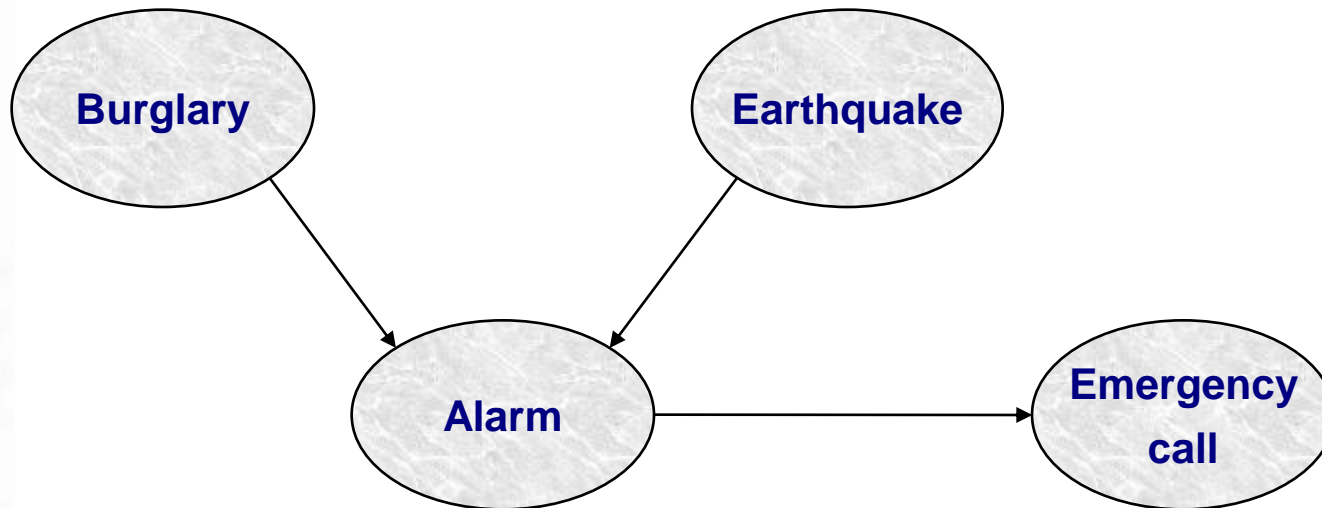




**Bayesian Networks procedure for creation  
of genes-biological issue dependence networks**

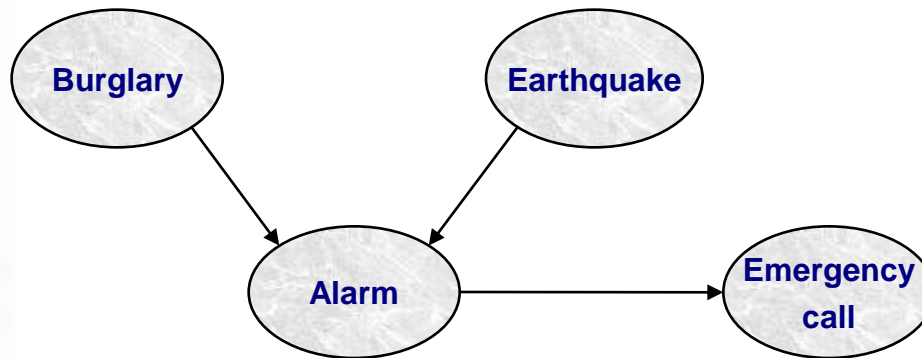


Bayesian networks provide us with a clear and simple tool for casual interactions representation





To fully describe Bayesian Network we need to know conditional distributions parameters



Probability of a burglary?

Probability of an earthquake?

Probability of running the alarm if no disaster happens?

Probability of running the alarm if there was only a burglary?

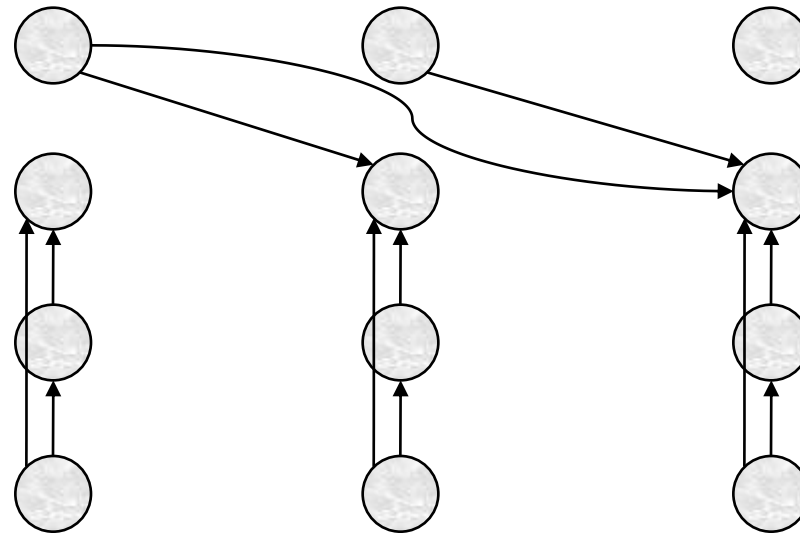
etc.

Issue: Is it a correct casual interaction graph?

The major problem is to find the proper graph that would describe the phenomenon under study



Dynamic Bayesian Networks



Data from microarray experiments may be time dependent

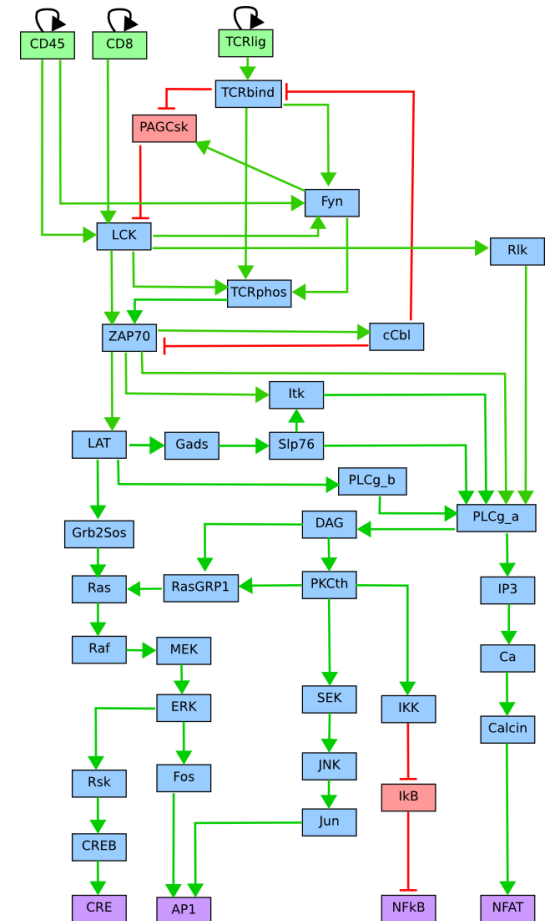
Ex. gene expression after 12, 24, 72 hours from the drug application



Given the data (obtained from NCBI) we are interested in finding the casual interactions network

The number of possible models grows at super-exponential rate ( $2^{n*n}$ ) when rising the number of nodes in the graph.

For microarray experiments the number of entities we deal with (genes, gene clusters, time step, disease stage, etc.) can reach thousands.



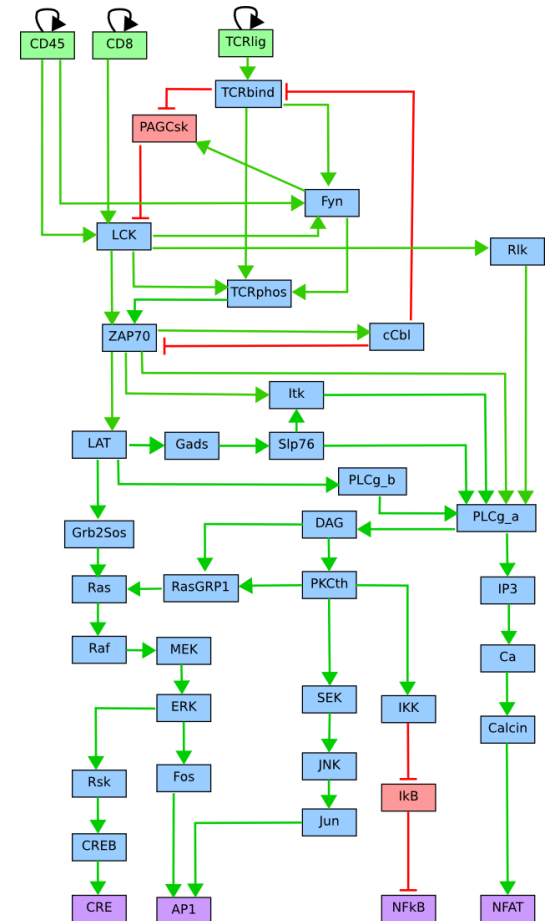


Possible solutions for the graphical model selection problem:

- ◆use heuristic methods which divide the problem into smaller ones
- ◆use **Markov Chain Monte Carlo simulation to draw samples from the posterior distribution**

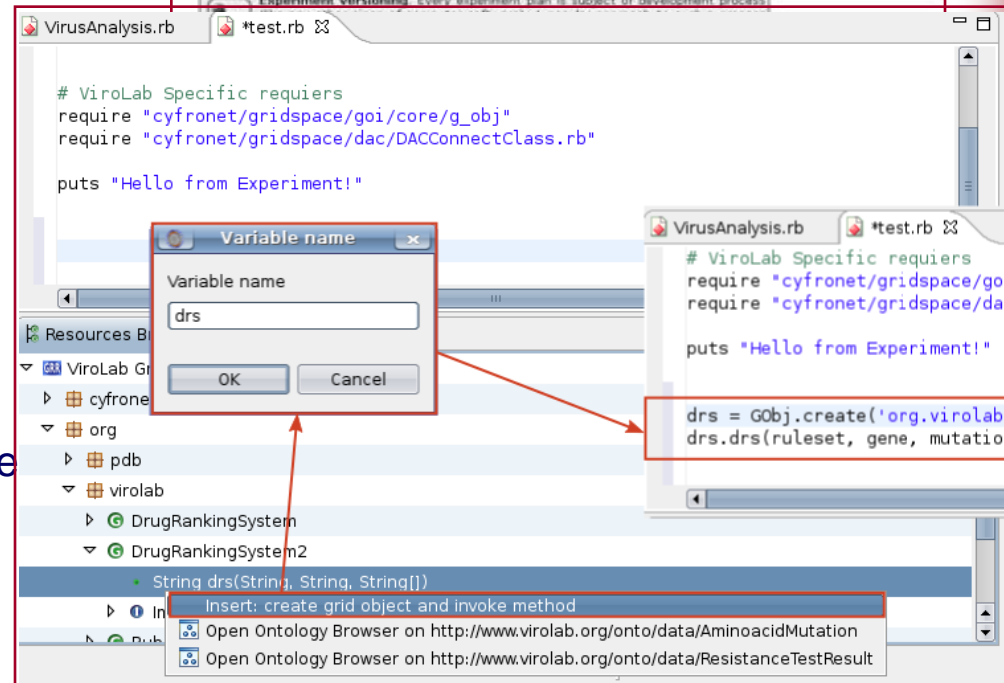
Calculation of the score function for the particular graph involves using complicated functions (such as the logarithm  $\Gamma$ -Euler function)

This together with overall complexity of the problem makes our task very computationally demanding



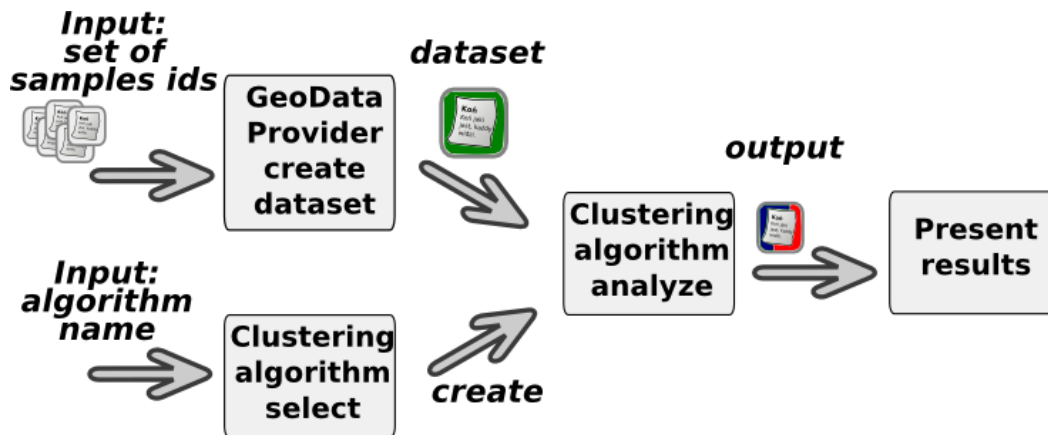


- Environment for development and execution of collaborative applications
- Scripting-based experiment plans (Ruby) for representing complex applications
- *Experiment Planning Environment* for experiment developers
- *Experiment Management Interface* (portal) for experiment users
- Experiment Repository, Result Management
- Access to wide range of middleware (*Grid, Web*)
- <http://virolab.cyfronet.pl>





- Structure of the tasks



- Available gems:

- DataProvider – retrieves data from NCBI, allows creating new datasets
- ClusteringAlgorithm – multiple algorithms are available (Agglomerative, Isodata, Shared Nearest Neighbor, Self-Organizing Maps, K-Means, Cobweb)
- ClusteringAlgorithmInstrumentation – Each algorithm may be tuned by setting different cluster metrics, sample metrics, cluster representation and cluster score functions



KU KDM 2009



Violab

Department of Bioinformatics and Telemedicine  
Jagiellonian University, Medical College



12-13.03.2009



- [NCBI-GEO]** Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar NCBI GEO: mining tens of millions of expression profiles—database and tools update *Nucleic Acids Research* Advance Access, 2007, DOI 10.1093/nar/gkl887. *Nucl. Acids Res.* 35: D760-D765.
- [Schafer]** Schafer J and Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21: 754-764.
- [Friedman]** Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science*, 30:799-805.
- [Seagal]** Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2): 166-76.
- [Murphy]** Murphy K and Mian S (1999) Modeling gene expression data using dynamic Bayesian networks. Technical Report, Computer Science Division, University of California, Berkeley, CA.
- [Fodor1993]** Fodor, S. P., et al. Multiplexed biochemical assays with biological chips. *Nature* 364, 555-6 (1993).
- [Kendall1979]** Kendall .M.G, Stuart A. (1979) *The advanced theory of statistics vol. 2.* 4th Ed. Griffin London.
- [Elston2002]** Elston R, Olson J, Palmer L. (2002) *Biostatistical genetics and genetic epidemiology.* John Wiley & Sons. Ltd. p 29-33 .
- [Piwowar2006]** Piwowar, M.; Meus, J.; Piwowar, P.; Wiśniowski, Z.; Stefaniak, J. & Roterman, I. (2006), 'Tandemly repeated trinucleotides - comparative analysis.', *Acta Biochim Pol* 53(2), 279—287
- [Milligan]** Glenn W. Glenn, Cooper C. Martha (1987) *Methodology Review: Clustering Methods.* Applied Psychological Measurement, Vol. 11, No. 4, 329-354